GenAIM: A Multimodal AI Music Generator with Only Lyrics or Images

Callie C. Liao
Stanford University
Stanford, CA USA
ccliao@stanford.edu

Ellie L. Zhang
IntelliSky
McLean, VA USA
elzhang@intellisky.org

Duoduo Liao
George Mason University
Fairfax, VA USA
dliao2@gmu.edu

Abstract

As generative Artificial Intelligence (AI) rises, AI for music generation has emerged as an increasingly prominent area of research. However, many existing neural-based music generation models heavily depend on large datasets, raising concerns about potential copyright infringement due to data training and increased costs to improve performance. In contrast, we propose Generate AI Music (GenAIM), an innovative multimodal AI music generation web tool for lyric-to-song, text-to-music, and image-to-music generation powered by a novel pure algorithm-driven music core. This music core incorporates both lyrical and non-lyrical inputs, such as text and images, and its pure algorithms overall effectively mitigate the risk of copyright infringement. The novel, purely algorithmic music generation process generates a coherent and flowing melody abiding music theory, lyrical, and rhythmic conventions. Users can generate music through the webpage nearly instantly. The webpage provides generated sheet music and the ability to play the music audio for listening. Overall, GenAIM can serve as a co-pilot tool to inspire current composers and lower the entry gap for musicians aspiring to transform thoughts to music by utilizing lyrics or visual imagery as a starting point. It also does not encroach upon human creativity, becoming a reliable music composition assistant and potential educational composition tutor. GenAIM is designed for individuals of all backgrounds, requiring no formal music training, and provides a range of benefits, including entertainment, relaxation, and support for mental well-being.

Click on the web link to run GenAIM to generate music.

1 Introduction

Artificial Intelligence Generated Content (AIGC) has grown rapidly, particularly with tools like ChatGPT Achiam et al. [2023]. However, AI music generation lags behind AI art and writing due to its complex structure and required musical expertise. Current methods rely on deep learning Agostinelli et al. [2023]Chen et al. [2024], using large datasets for music generation, but face challenges including data collection, copyright risks, high computing costs, and labor-intensive data preparation Bao et al. [2019]Copet et al. [2023]Dong and Yang [2018].

In developing a musically robust song generation method, knowledge of music theory, literature, and linguistics is required to emulate the musical and lyrical intuitive thinking process without relying on existing music. For this pilot study, we encompass the relationship between lyrics and melody as well as music theory guidelines to create Generate AI Music (GenAIM), a lyric-to-song, image-to-music, and text-to-music generation web tool powered by a novel pure algorithm-driven music core. Only the image-to-music pipeline Zhang et al. [2025] employs Large Language Models

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The First Workshop on Generative and Protective AI for Content Creation.

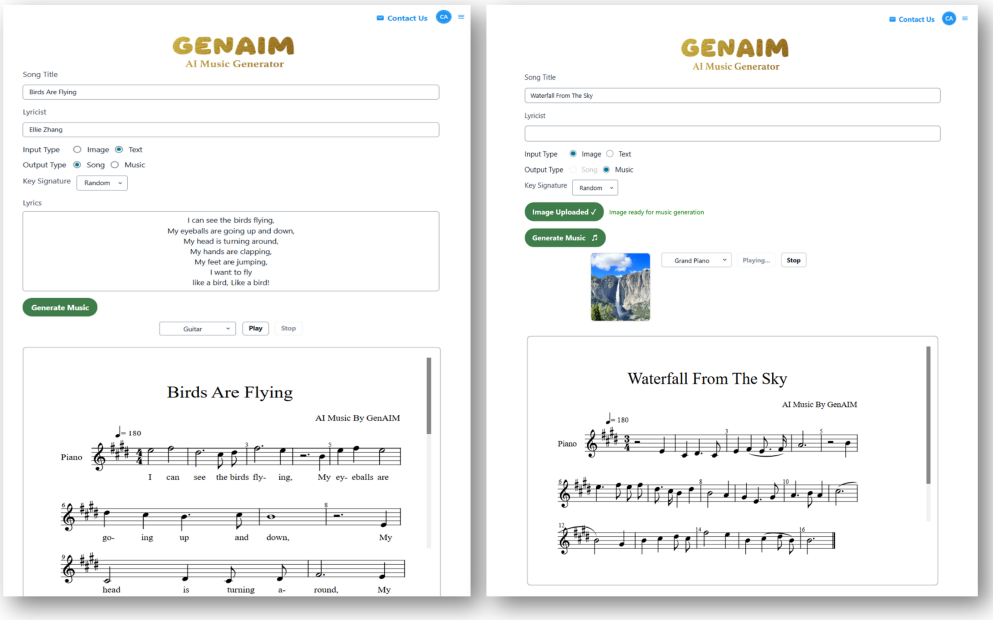


Figure 1: Music generation from lyrics (sample lyrics by Zhang [2016]) and the image (sample images by the authors).

(LLMs) to produce lyrical content for music generation because we leverage lyric-music correlations in the music core to generate music, as inspired by Liao et al. [2022]Liao et al. [2024]. Figure 1 presents two representative examples that demonstrate the system's capability to process both textual and visual modalities: a text-based input and an image-based input. Our generation method also includes features such as customizable key signatures and instruments for playback, sheet music for display, and the ability to view the user's music history as shown in Figure 2. The novelty of our non-deep learning lyric-to-music algorithms for the music core within GenAIM, along with the facilitation of transforming multimodal thoughts to music while preserving human creativity, are key contributions of our GenAIM web tool, drastically differing from existing music generation models.

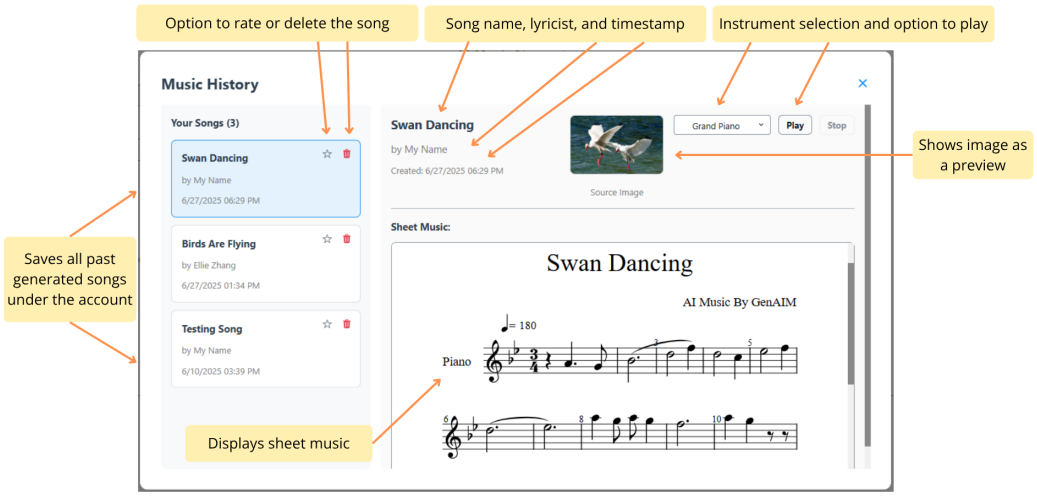


Figure 2: The demo music history feature.

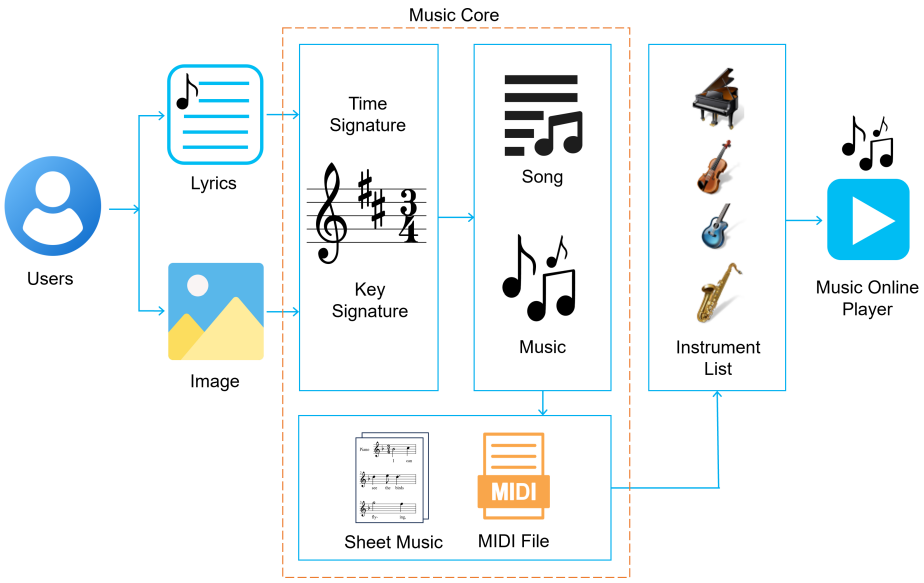


Figure 3: The system architecture of GenAIM.

2 The System Architecture

Figure 3 displays the system architecture of GenAIM. This system is built on the Amazon Web Services (AWS) platform. The input image is processed by LLMs provided by the AI services, which generate the lyrics. These generated lyrics are then passed to the music core for composition. The front-end web application loads the music files from the AWS cloud, renders them to sheet music, and plays the music.

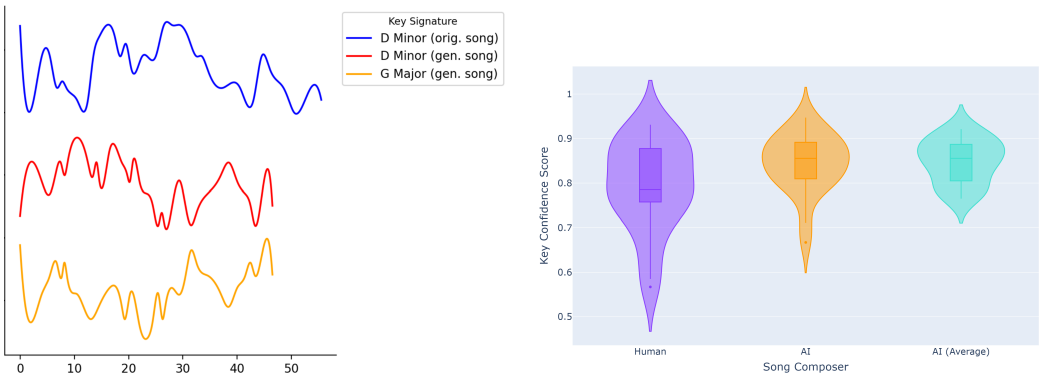
At the user level, two forms of input are accepted: lyrics and image. Users can also select their preferred key signature or the "random" option that at least selects a moderately fitting key, while the most fitting time signature is determined based on the input lyrics by the algorithm itself as described in Liao et al. [2023]. The rhythmic structure is essentially defined by the time signature and lyrical keywords associated with strong beats as identified in Liao et al. [2022]Liao et al. [2024].

The generation options available to the users are non-lyrical music generation or lyrical song generation. Inside the system, an input image or a set of lyrics is transferred to the music core, where pitches, rhythms, phrases, etc. are formed based on the input as detailed in Zhang et al. [2025]Liao et al. [2025]. After constructing the score and pitch, the music is converted into a MusicXML file for better readability and compatibility with composition and notation software, facilitating digital sheet music exchange and collaboration. The MusicXML file can also be converted again to MIDI format or displayed as sheet music.

When the music result has been generated, a dropdown menu offers a variety of instruments that the user can select for playback, such as the guitar and piano. A feature is also provided to access all previously generated music as shown in Figure 2, where users can rate or delete saved songs, select an instrument for playback, and replay the song. Additional information such as sheet music, song title, lyricist, song generation timestamp, and the image, if applicable, are displayed as well. Overall, this feature is intuitive and easy to navigate.

3 Experiments and Evaluations

As part of the experiments for this demonstration system, we performed exploratory data analysis using Music21 [Cuthbert and Ariza, 2010–], a Python-based development toolkit created by the Massachusetts Institute of Technology (MIT), to identify similarities and differences between AI-generated and original compositions. To evaluate the generated music in comparison to human-composed pieces, 25 original lyrics from piano books are selected to generate 4 to 5 songs per lyrical



(a) Melody motion comparisons.

(b) Key confidence with various key signatures.

Figure 4: AI-generated songs vs. original songs

song group in various keys, resulting in 112 AI-generated pieces. Each set includes at least one version matching the time and key signature of the original song, enabling direct comparison for structural and musical analysis.

GenAIM is capable of generating melodies resembling human-composed melodies. Figure 4a presents a comparison of melodic motion between the original and AI-generated songs. The generated song in the same key as the original exhibits a highly similar motion with the original song, and the generated song in a different key displays variation from the original but retains smoothness. These comparisons demonstrate the reliance of GenAIM in melody generation.

Figure 4b shows a violin plot comparing key confidence scores derived from music21 algorithms, across human composers, an AI composer, and an averaged AI composer. The AI composer retains the original key and time signatures, while the averaged AI composer generates songs without those constraints. Human compositions show a broader distribution, whereas AI composers produce more concentrated scores, particularly in the 0.8–0.9 range, suggesting greater consistency and alignment with a well-defined tonal center in AI-generated music.

4 Conclusions & Future Work

We introduced GenAIM, a web tool powered by an algorithm-driven music core. GenAIM enables music generation by integrating both lyrical and non-lyrical inputs, including text and images. Furthermore, the lyric-to-music generation process does not rely on prior training data. Overall, GenAIM is capable of generating highly fitting melodies rhythmically and melodically, and our novel method can inspire music generation techniques from a refreshing perspective. As our web tool evolves, our method will benefit both amateur and professional musicians as a copilot and an educational tutor, accelerating the composition workflow and learning process for aspiring musicians, while fostering better mental well-being and entertainment for individuals of all backgrounds.

Since this pilot project demonstration is solely focused on the generation of main melodies, no further musical complexity is incorporated. For future work, the pitch construction can still be improved upon, such as including chord progressions and cadences. Furthermore, more diverse experiments with different genres can be considered.

5 Ethical Implications

AI music generation with deep learning poses ethical challenges. The proposed methods in this paper does *not* use deep learning in the text-to-music process, but LLMs are utilized to process the image for music generation. Nonetheless, LLMs are only used to capture the features of the images and are ultimately used solely to help form the rhythmic structure and inspire the music; thus, there is no direct use of LLMs in the music generation pipeline, avoiding copyright infringement issues and presenting the AI-generated songs as completely original work. In addition, in this paper, all the sample images, text, and lyrics were taken or created by the authors, so there are no copyright issues.

References

J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *ArXiv Preprint ArXiv:2303.08774*, 2023.

A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

H. Bao et al. Neural melody composition from lyrics. *NLPCC*, 11838:499–511, 2019.

K. Chen, Y. Wu, H. Liu, M. Nezhurina, T. Berg-Kirkpatrick, and S. Dubnov. Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2024.

J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Defossez. Simple and controllable music generation. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 47704–47720. Curran Associates, Inc., 2023.

M. S. Cuthbert and C. Ariza. music21: A toolkit for computer-aided musicology. `https://web.mit.edu/music21/`, 2010–.

H.-W. Dong and Y.-H. Yang. Convolutional generative adversarial networks with binary neurons for polyphonic music generation. *arXiv preprint arXiv:1804.09399*, 2018.

C. C. Liao, D. Liao, and J. Guessford. Multimodal lyrics-rhythm matching. In *Proc. of the 2022 IEEE Int. Conf. on Big Data (BigData)*, pages 3622–3630, 2022. doi: 10.1109/BigData55660.2022.10021009.

C. C. Liao, D. Liao, and J. Guessford. Automatic time signature determination for new scores using lyrics for latent rhythmic structure. In *Proc. of the 2023 IEEE Int. Conf. on Big Data (BigData)*, pages 4485–4494, 2023. doi: doi.org/10.1109/BigData59044.2023.10386875.

C. C. Liao, D. Liao, and E. L. Zhang. Relationships between Keywords and Strong Beats in Lyrical Music . In *2024 IEEE International Conference on Big Data (BigData)*, pages 3191–3199, Los Alamitos, CA, USA, Dec. 2024. IEEE Computer Society. doi: 10.1109/BigData62323.2024.10825973. URL `https://doi.ieeecomputersociety.org/10.1109/BigData62323.2024.10825973`.

C. C. Liao, D. Liao, and E. L. Zhang. MusicAIR: A multimodal AI music generation framework powered by an algorithm-driven core. In *2025 IEEE International Conference on Big Data (BigData)*, Los Alamitos, CA, USA, Dec. 2025. IEEE Computer Society.

E. L. Zhang. Birds are flying. *Journal of Children's Music*, 365:9, Nov. 2016.

E. L. Zhang, C. C. Liao, and D. Liao. IMA$_i$Gen: A cross-modal image-to-music generation with a non-deep learning core. In *2025 IEEE MIT Undergraduate Research Technology Conference (URTC)*, 2025.